

Electronic Supplementary Material

Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta, and Cryptista

Fabien Burki, Maia Kaplan, Denis V. Tikhonov, Vasily Zlatogursky, Bui Quang Minh,
Liudmila V. Radaykina, Alexey Smirnov, Alexander P. Mylnikov, Patrick J. Keeling

Supplemental Material and Methods

Culturing

Clone HF-7 (*Choanocystis* sp.) was obtained from the wastewater ditch of treatment facilities near the settlement of Borok, Yaroslavskaya oblast, Russia (58.070 N, 38.238 E) on March 20, 2013. The water sample was collected at 20 cm depth and contained mainly organic detritus. Clone HF-20 (*Acanthocystis* sp.) was obtained from the upper reaches of the Zavkhan River, Western Mongolia, on August 2012. The sample was collected in the littoral of upper reaches of Taishir reservoir. Water sample containing *Raphidiophrys heterophryoidea* Zlatogursky 2012 (clone 00434) was collected on July 23, 2012 from Lake Nikonovskoe (*Raphidiophrys heterophryoidea* Zlatogursky 2012). The sample containing *Raineriophrys erinaceoides* [1](clone 00344) was taken on August 4, 2010 from Lake Leshevoe. Both lakes are located on Valamo Island, Lake Ladoga, North-Western Russia (61.383 N, 30.9 E). Clonal cultures were isolated from single cells using a micromanipulator fitted with a glass micropipette. Single cells were transferred to a Petri dish containing a clonal culture of the bacteriotrophic flagellate *Bodo saltans* Ehrenberg, 1832 as food. *B. saltans* (strain PlF-2, collection of live protozoan cultures at Institute of the Biology of Inland Waters, Russian Academy of Sciences, Russia (IBIW RAS)) was cultivated in the modified Pratt medium ($\text{KNO}_3 - 100 \text{ mg l}^{-1}$; $\text{K}_2\text{HPO}_4 - 10 \text{ mg l}^{-1}$; $\text{MgSO}_4 \cdot 7\text{H}_2\text{O} - 10 \text{ mg l}^{-1}$; $\text{FeCl}_3 \cdot 6\text{H}_2\text{O} - 1 \text{ mg l}^{-1}$) or Prescott-James medium ($\text{CaCl}_2 \cdot 2\text{H}_2\text{O} - 43 \text{ mg l}^{-1}$; $\text{KCl} - 16 \text{ mg l}^{-1}$; $\text{K}_2\text{HPO}_4 - 51 \text{ mg l}^{-1}$; $\text{MgSO}_4 \cdot 7\text{H}_2\text{O} - 28 \text{ mg l}^{-1}$) with addition of *Pseudomonas fluorescens* bacteria as food. The clones HF-20 and HF-7 are stored in the collection of live protozoan cultures at IBIW RAS.

New data from an unpublished transcriptomic dataset for an undescribed Apusomonads species were also included (Table S3); clone AF-17 (*Amastigomonas* sp.) was isolated from a soil and mosses sample collected on Mo Hill (Mo Shan), East Lake Area, Wuhan, China on March 22, 2012. Clonal culture was obtained from a single flagellated cell. RNA extraction and cDNA synthesis procedures were the same as for the centrohelids (see below).

cDNA preparation

Cells grown in clonal laboratory cultures were harvested following peak abundance after eating most of the prey. Cells were collected by centrifugation (2000 x g, room temperature) on the 0.8 µm membrane of Vivaclear Mini columns (Sartorium Stedim Biotech Gmng, Germany, Cat. No VK01P042) in the case of *Acanthocystis* sp. and *Choanocystis* sp., or by manual picking of about 2000 cells in the case of *R. heterophryoidea* and *R. erinaceoides*. Total RNA was extracted using the RNAAqueous kit (Ambion, lot # 1304062) and was converted into cDNA prior to sequencing using the SMARTer technology (SMARTer Pico PCR cDNA Synthesis Kit, Clontech, lot # 1308018A).

Sequencing and assembling

Sequencing was performed on the Illumina MiSeq platform with read lengths of 250bp, using the NexteraXT protocol to construct paired-end libraries. Read quality was assessed with FastQC [2] before and after quality trimming and SMART adaptors removal, which was performed with FastqMcf [3]. Cleaned reads were assembled into contigs with Trinity r20140717 using default parameters, followed by open reading frame (ORF) extraction using TransDecoder (part of the Trinity package). To assess the proportion of food sequences and other contamination in our data, the Blobology pipeline [4] was used against a custom database consisting of NCBI NT + *Bodo saltans* genome (food source). This protocol did not identify any large-scale contamination, but confirmed the presence of very little *B. saltans* sequences, ($\leq 1\%$) as well as human ($\leq 0.4\%$) and prokaryotic sequences ($\leq 0.2\%$). *Choanocystis* sp. was the least clean dataset, with $\leq 1.6\%$ of sequences deemed contaminants; the other 3 datasets generally contained $\leq 0.1\%$ of total foreign sequences. In an attempt to automatically remove eukaryotic contamination (prokaryotes were ignored due to very low level and not to remove potentially important organelle-related genes), the centrohelid contigs were blasted separately against *B. saltans* and *human* genomes, and the hit similarity sorted according to their frequency distributions. Based on these distributions, the contigs displaying similarities $\geq 80\%$ to either *B. saltans* or *human* were discarded. Finally, to prepare the contigs for phylogenomic dataset construction, a filtering step was applied on the translated sequences to remove potential untranslated regions (UTR) or short artifactual sequences resulting from poor quality assembling. To do so, the protein sequences were searched by BLASTP against the highly curated SwissProt database, and only the segments producing significant alignments were retained (evalue cutoff $\leq 1e-5$).

Single-genes preparation

Following the removal of putative contaminants, all remaining centrohelid contigs were searched for a set of 263 genes (referred hereafter as seed genes) previously used in phylogenomic analyses [5-7]. First, all sequences already included in the seed genes, representing a good eukaryotic diversity, were used as queries in separate BLASTP searches against the new contigs of each species. The top 4 hits were retrieved in non-redundant manner and appended to the seed genes. The same procedure was applied to 127 newly available protein collections predicted from transcriptomic and genomic datasets, mostly released by the MMETSP consortium [8]; see Table S3 for the complete list of new taxa and references. Of particular notes were i) the replacement of genome data of *Emiliania huxleyi* [9] by new transcriptome sequences after noticing many ambiguities stemming from gene modeling inaccuracy, and ii) all new MMETSP ciliate data were retranslated with the EMBOSS Transeq tool using the ciliate codon table due to the presence of large frameshifts in the available predicted sequences. The expanded seed genes were then aligned with MAFFT-LINSI v.7, and ambiguously aligned positions were trimmed off with TRIMAL v1.4 (maximum number of gaps allowed per site: 20%). A final check was performed on all multiple sequence alignments (MSA) by visual inspection using AliView [10]. To identify contaminants and all homologous copies introduced by selecting the top 4 blast hits, individual genes were subjected to Maximum Likelihood (ML) tree reconstructions, using RAxML v.8 in combination with 100 rapid bootstraps and the LG empirical rate matrix with four gamma categories for handling the rate heterogeneity across sites and amino acid frequencies computed from the data (LG+Γ4+F model). Single-gene trees were manually screened by independent inspections (two researchers) to flag and discard from further analysis contaminating and paralogous sequences.

Supplemental references

1. Mikrjukov, K. A. 2001 Heliozoa as a component of marine microbenthos: A study of heliozoa of the White Sea. *Ophelia* **54**, 51–73.
2. Andrews, S. 2010 FastQC: a quality control tool for high throughput sequence data.
3. Aronesty, E. 2013 Comparison of Sequencing Utility Programs. *Expression Analysis*, Durham, NC **7**, 1–8.
4. Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M. & Blaxter, M. 2013 Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front Genet* **4**, 237.
5. Burki, F., Okamoto, N., Pombert, J.-F. & Keeling, P. J. 2012 The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. *P R Soc B* **279**, 2246–2254.
6. Burki, F., Corradi, N., Sierra, R., Pawlowski, J., Meyer, G. R., Abbott, C. L. & Keeling, P. J. 2013 Phylogenomics of the intracellular parasite *Mikrocytos mackini* reveals evidence for a mitosome in rhizaria. *Curr Biol* **23**, 1541–1547.
7. Janouškovec, J., Tikhonenkov, D. V., Burki, F., Howe, A. T., Kolisko, M., Mylnikov, A. P. & Keeling, P. J. 2015 Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. *Proc Natl Acad Sci U.S.A.*, 201423790–8.
8. Keeling, P. J. et al. 2014 The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol* **12**, e1001889.
9. Read, B. A. et al. 2013 Pan genome of the phytoplankton *Emiliania* underpins its global distribution. **499**, 209–213.
10. Larsson, A. 2014 AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278.

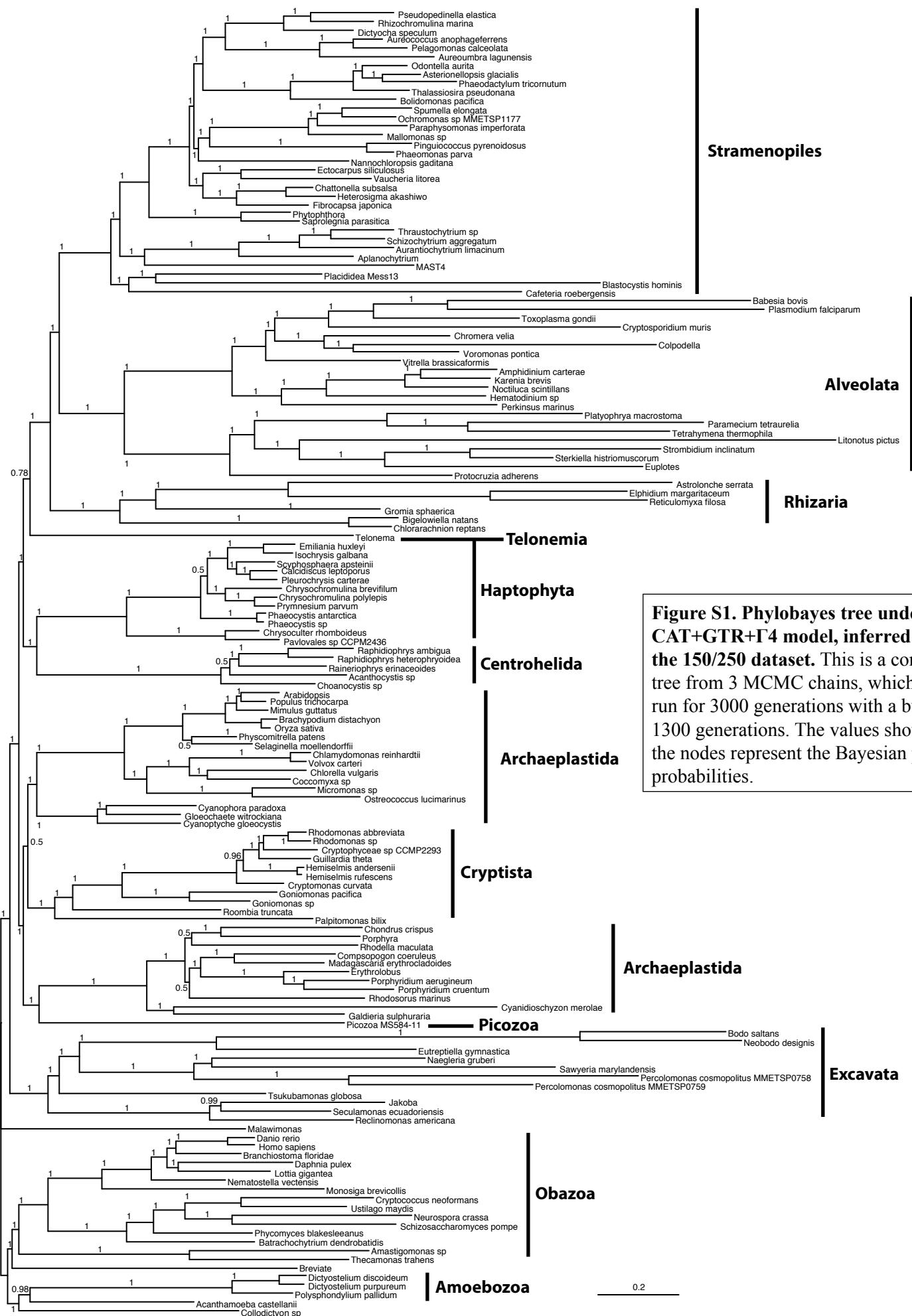


Figure S1. Phylobayes tree under the CAT+GTR+ Γ 4 model, inferred from the 150/250 dataset. This is a consensus tree from 3 MCMC chains, which were run for 3000 generations with a burnin of 1300 generations. The values shown at the nodes represent the Bayesian posterior probabilities.

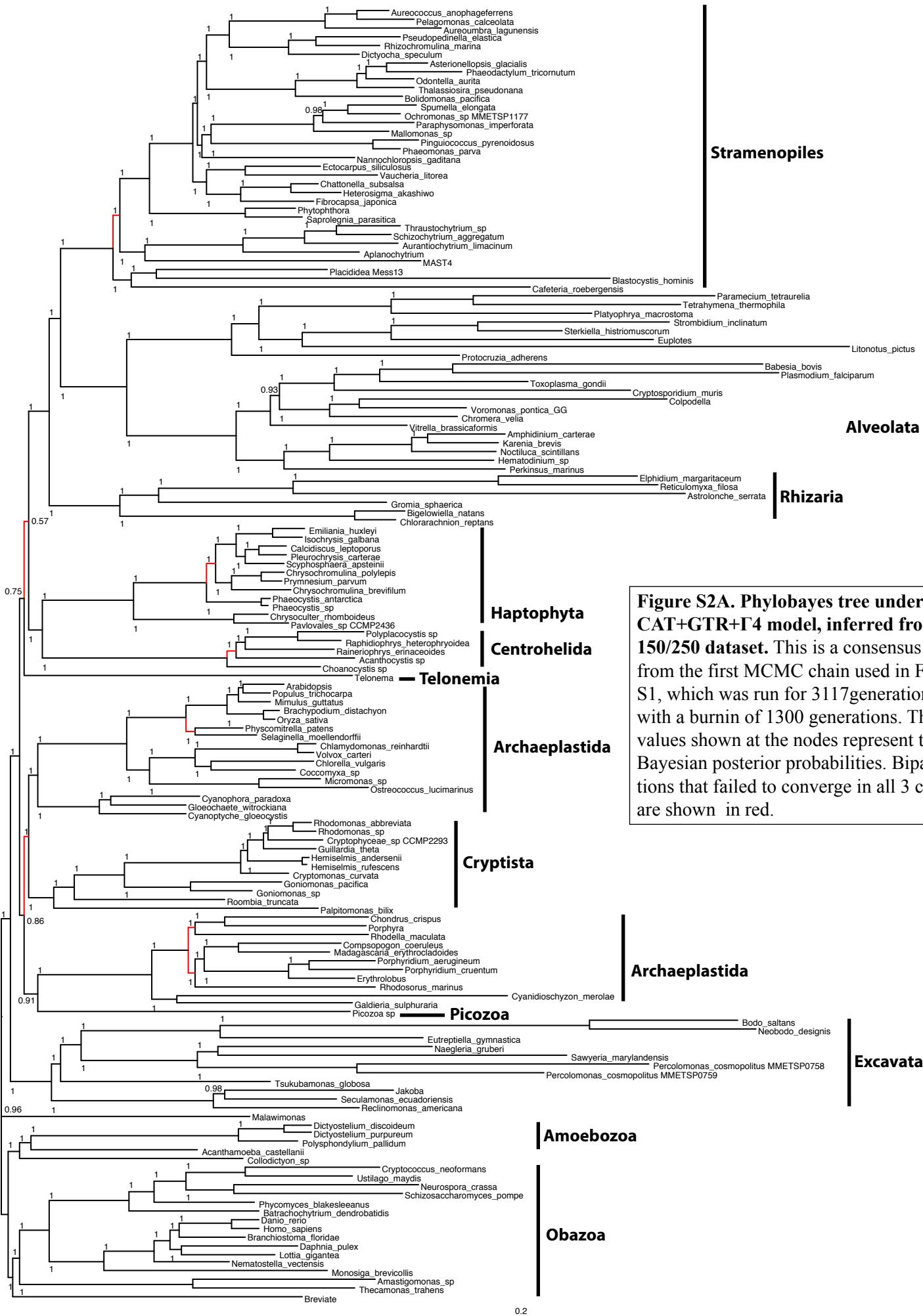


Figure S2A. Phylobayes tree under the CAT+GTR+ Γ 4 model, inferred from the 150/250 dataset. This is a consensus tree from the first MCMC chain used in Figure S1, which was run for 3117 generations with a burnin of 1300 generations. The values shown at the nodes represent the Bayesian posterior probabilities. Bipartitions that failed to converge in all 3 chains are shown in red.

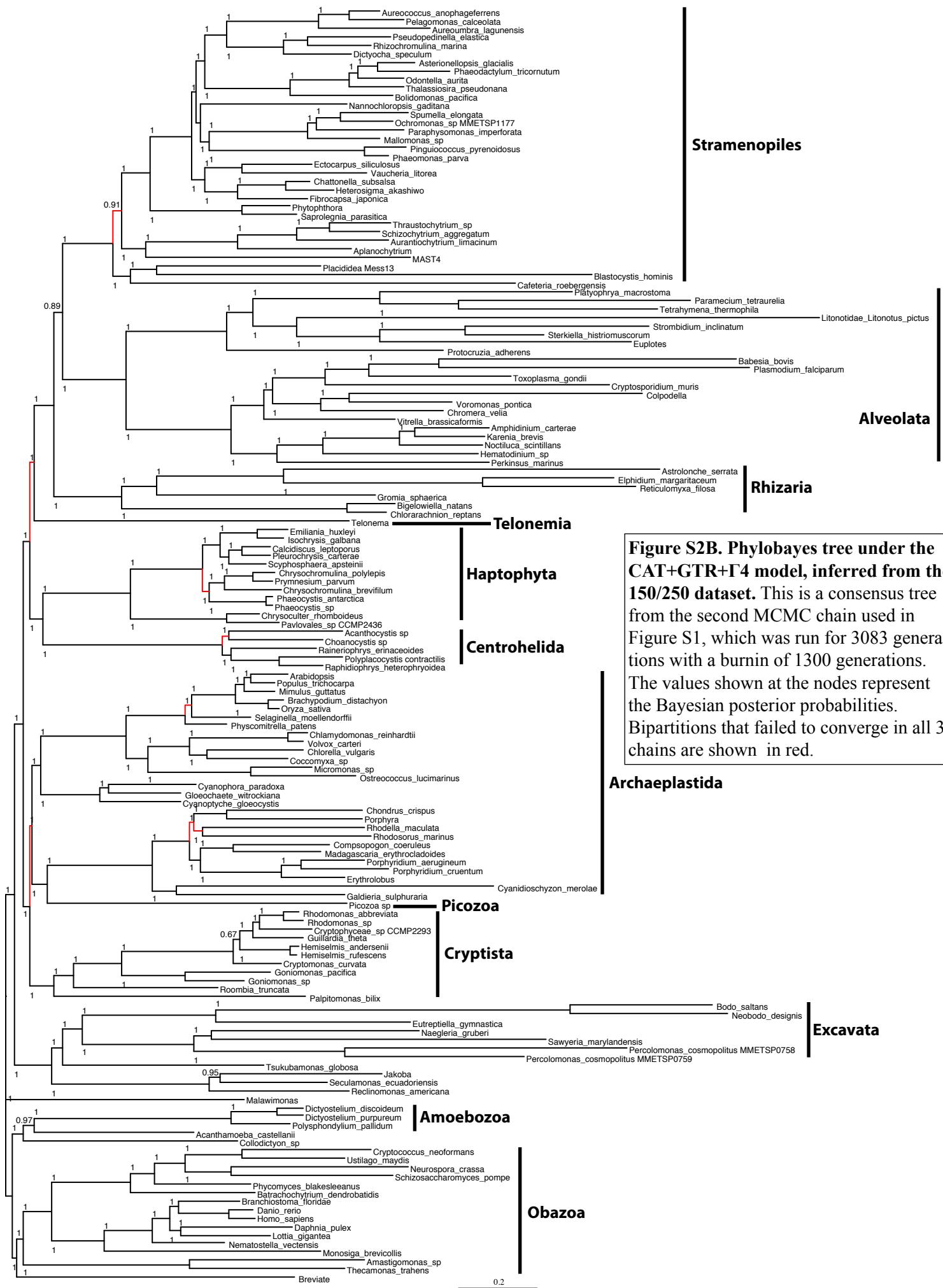


Figure S2B. Phylobayes tree under the CAT+GTR+ Γ 4 model, inferred from the 150/250 dataset. This is a consensus tree from the second MCMC chain used in Figure S1, which was run for 3083 generations with a burnin of 1300 generations. The values shown at the nodes represent the Bayesian posterior probabilities. Bipartitions that failed to converge in all 3 chains are shown in red.

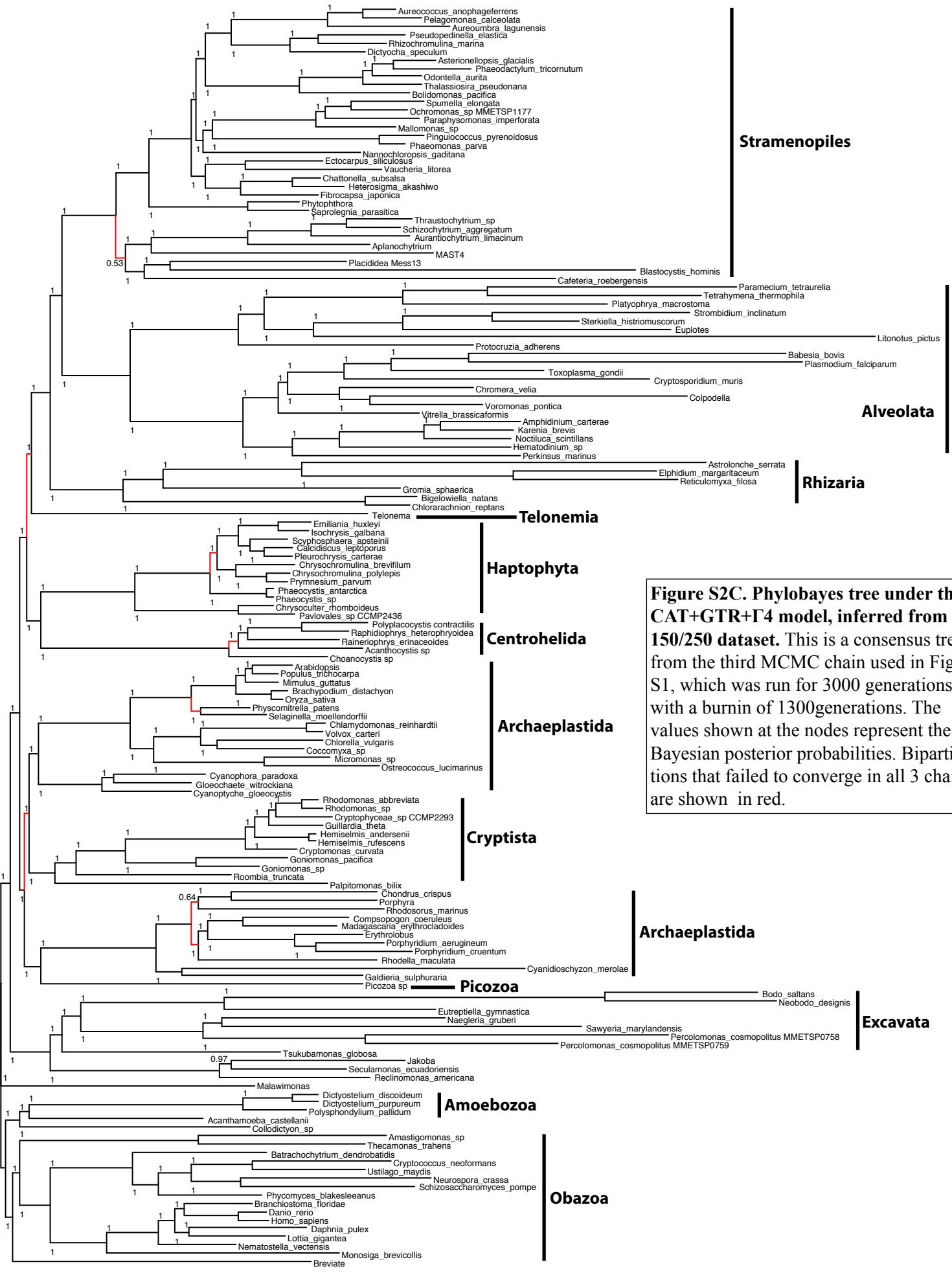


Figure S2C. Phylobayes tree under the CAT+GTR+G4 model, inferred from the 150/250 dataset. This is a consensus tree from the third MCMC chain used in Figure S1, which was run for 3000 generations with a burnin of 1300 generations. The values shown at the nodes represent the Bayesian posterior probabilities. Bipartitions that failed to converge in all 3 chains are shown in red.

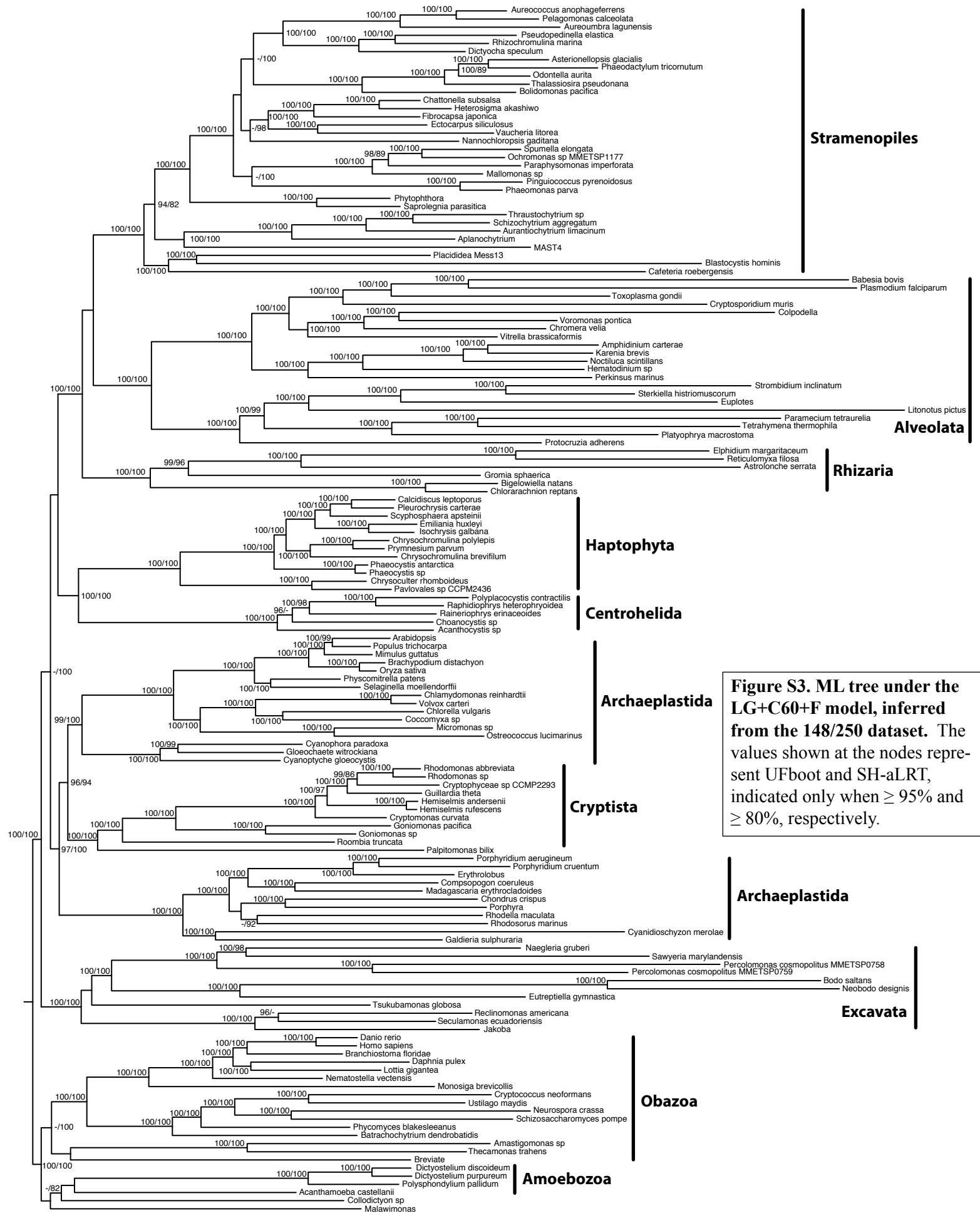


Figure S3. ML tree under the LG+C60+F model, inferred from the 148/250 dataset. The values shown at the nodes represent UFboot and SH-alRT, indicated only when $\geq 95\%$ and $\geq 80\%$, respectively.

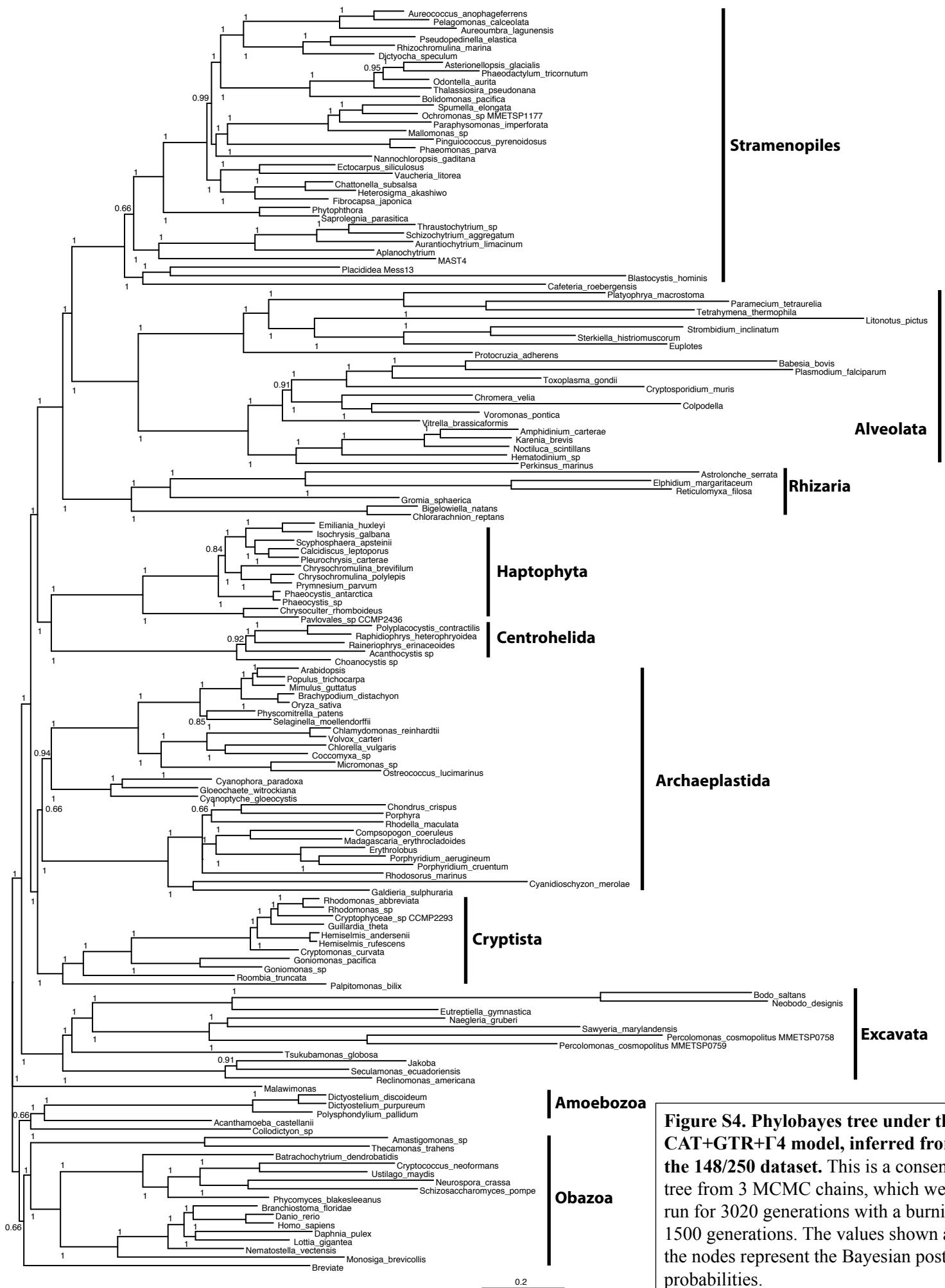


Figure S4. Phylobayes tree under the CAT+GTR+Γ4 model, inferred from the 148/250 dataset. This is a consensus tree from 3 MCMC chains, which were run for 3020 generations with a burnin of 1500 generations. The values shown at the nodes represent the Bayesian posterior probabilities.

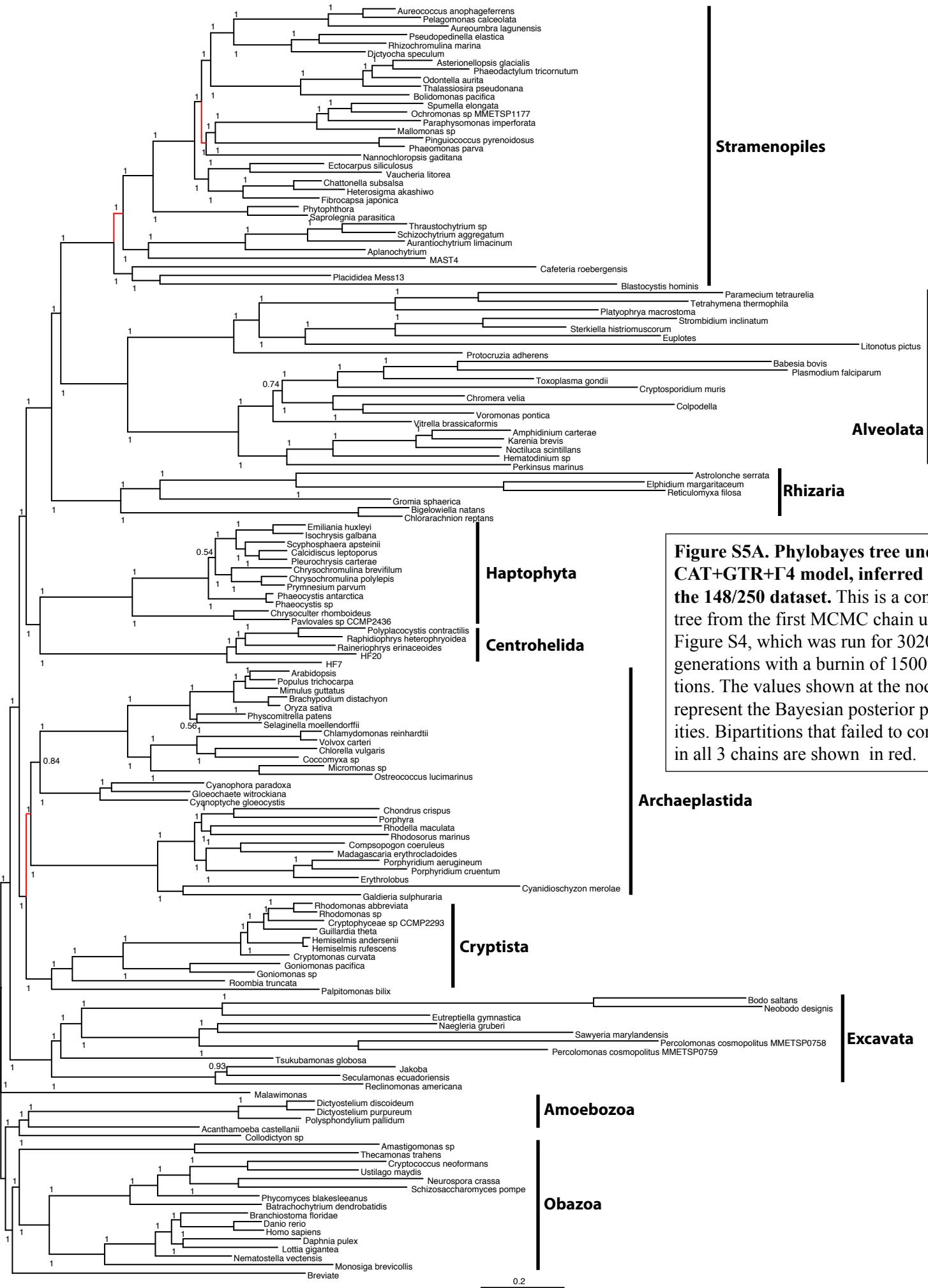


Figure S5A. Phylobayes tree under the CAT+GTR+G4 model, inferred from the 148/250 dataset. This is a consensus tree from the first MCMC chain used in Figure S4, which was run for 3020 generations with a burnin of 1500 generations. The values shown at the nodes represent the Bayesian posterior probabilities. Bipartitions that failed to converge in all 3 chains are shown in red.

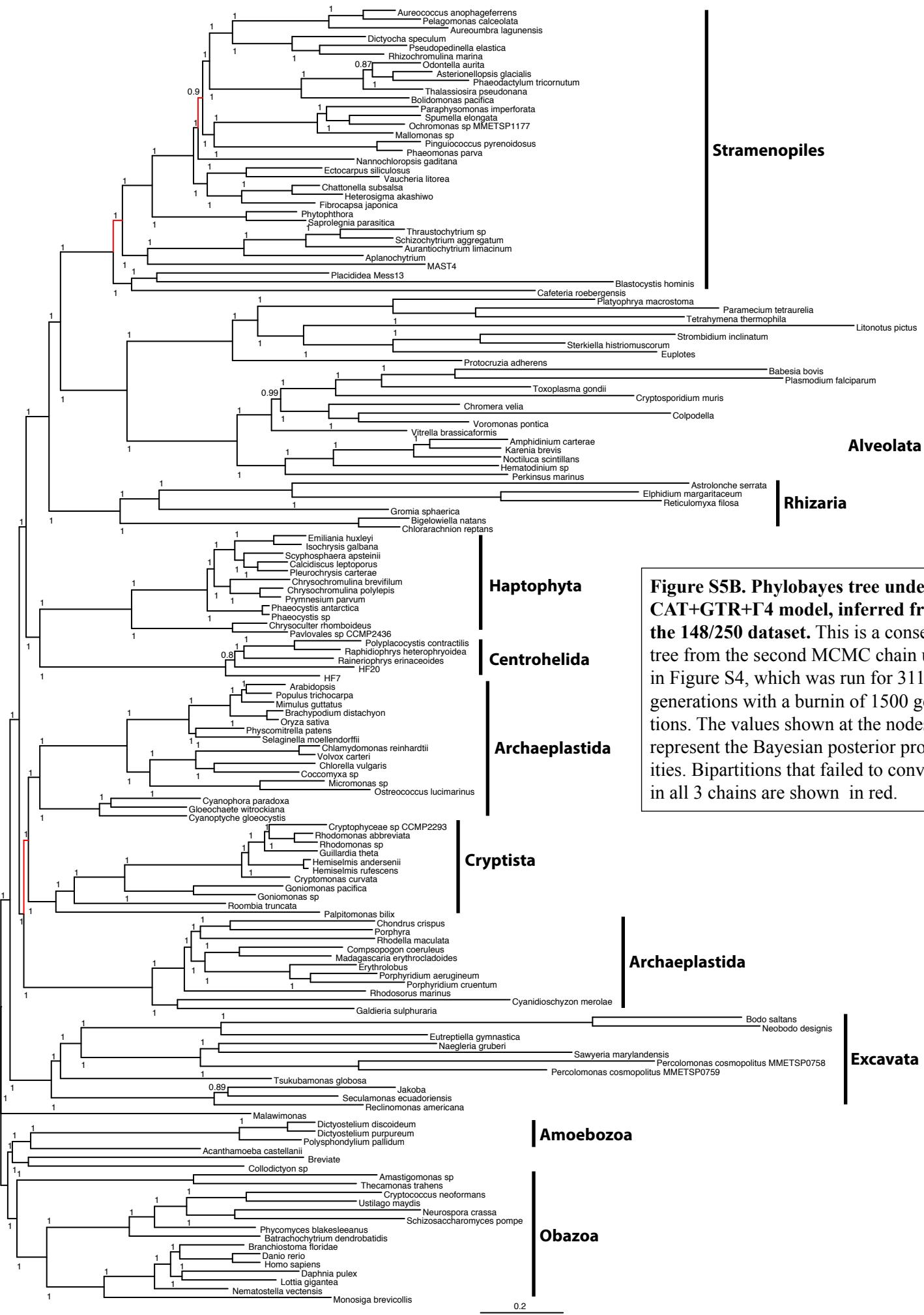


Figure S5B. Phylobayes tree under the CAT+GTR+G4 model, inferred from the 148/250 dataset. This is a consensus tree from the second MCMC chain used in Figure S4, which was run for 3117 generations with a burnin of 1500 generations. The values shown at the nodes represent the Bayesian posterior probabilities. Bipartitions that failed to converge in all 3 chains are shown in red.

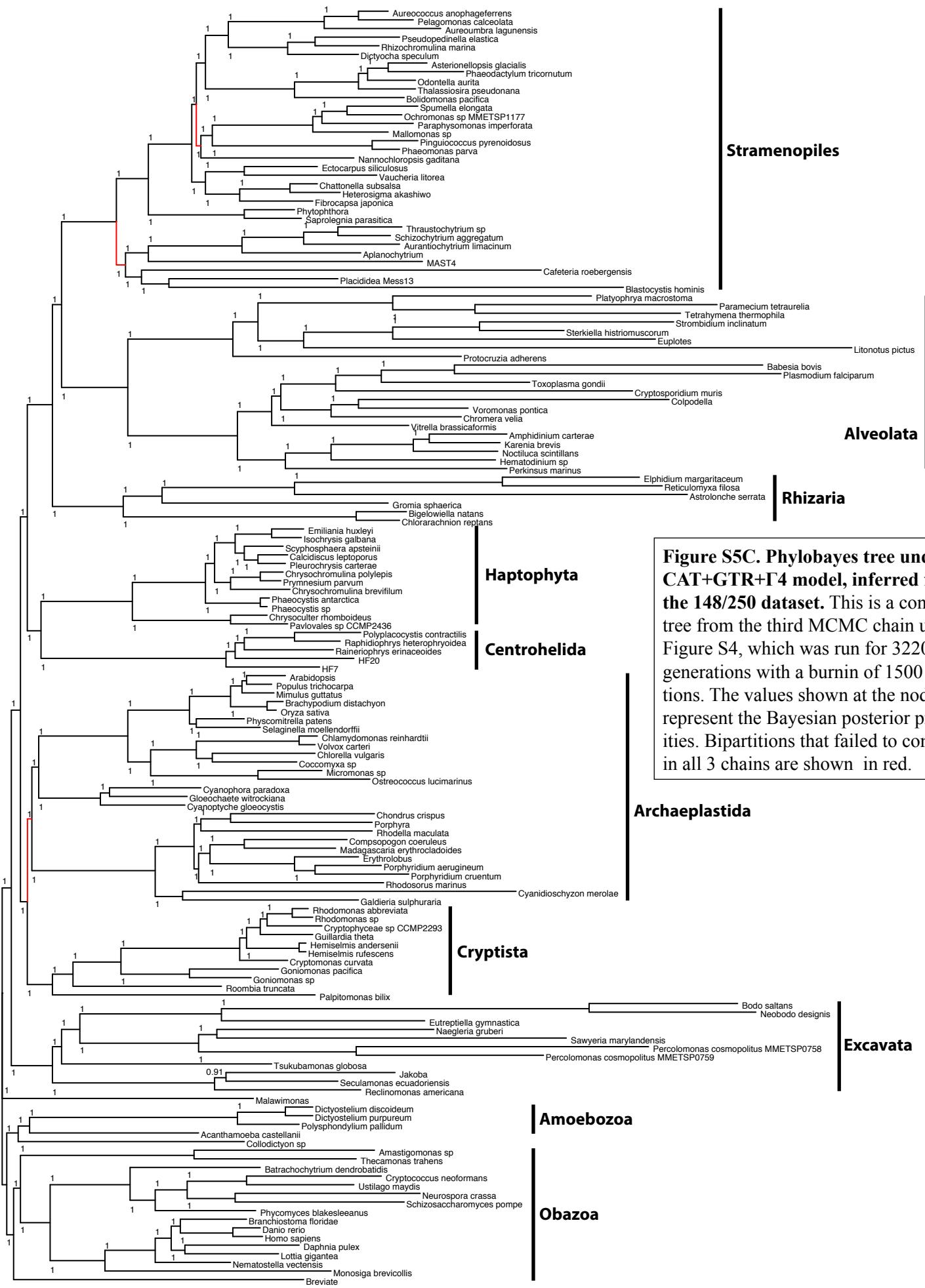


Figure S5C. Phylobayes tree under the CAT+GTR+ Γ 4 model, inferred from the 148/250 dataset. This is a consensus tree from the third MCMC chain used in Figure S4, which was run for 3220 generations with a burnin of 1500 generations. The values shown at the nodes represent the Bayesian posterior probabilities. Bipartitions that failed to converge in all 3 chains are shown in red.

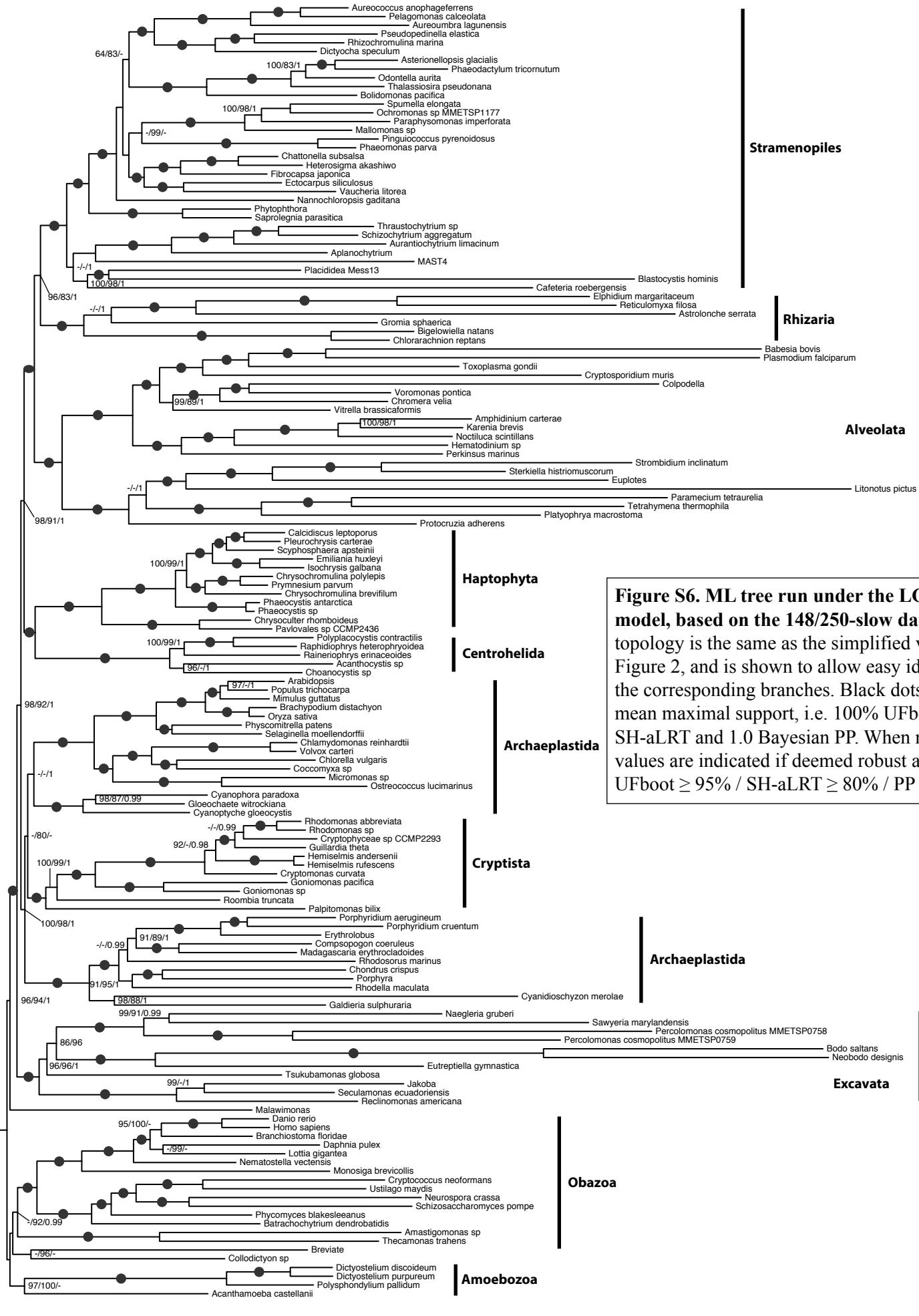


Figure S6. ML tree run under the LG+C60+F model, based on the 148/250-slow dataset. The topology is the same as the simplified version in Figure 2, and is shown to allow easy identification of the corresponding branches. Black dots on branches mean maximal support, i.e. 100% UFboot and SH-aLRT and 1.0 Bayesian PP. When not maximal values are indicated if deemed robust as followed: UFboot \geq 95% / SH-aLRT \geq 80% / PP \geq 0.99.